# DIFFERENCES BETWEEN THE METHOD OF MINI-MODELS AND THE K-NEAREST NEIGHBORS ON EXAMPLE OF MODELING OF UNEMPLOYMENT RATE IN POLAND

**Andrzej Piegat[a), b)], Barbara Wąsikowska[b)], Marcin Korzeń[a)]**

[a)] Faculty of Computer Science and Information Systems, Westpomeranian University of Technology, 71-210 Szczecin, Zolnierska 49, Poland
[b)]Faculty of Economics and Management, University of Szczecin, 71-101 Szczecin, Mickiewicza 64, Poland

**Abstract.** The paper presents in a possibly reader-friendly way, in the 2D-space, the method of mini-models, which suits very good for modeling economic dependencies, where frequently a part of explanatory variables influencing the resulting variable is not known (lack of data). Experiments realized by authors confirmed superiority of mini-models over such modeling methods as polynomials, neural network, and the method of K-nearest neighbors (KNN). Because the method of mini-models is frequently mistaken for the KNN-method the authors explain in the paper the significant difference between the both competitive methods. The indicated difference is also the main reason of superiority of mini-models over the KNN-method. Accuracy of both methods has been compared experimentally on example of modeling unemployment rate in Poland and also on examples of other economic dependencies.

**Keywords.** Quantitative methods in data analyses and modeling, modeling of business processes.

## 1. INTRODUCTION

The concept and formulas for mini-models have been developed by Andrzej Piegat [6] and experiments were carried out by Barbara Wąsikowska and Marcin Korzeń. Method of mini-models serves for calculation of numerical answers (predictions) for numerical questions concerning the explained variable y for a given vector of explanatory variables $x_1, x_2, ..., x_n$. Thus, the task of a mini-model is the same as of a usual global model of the dependence $y=f(x_1, x_2, ..., x_n)$. The difference consists in the fact that the global model encircles the full space $X_1*X_2*...*X_n$ of the explanatory variables and mini-model encircles only a certain local part of the full space, however this part in which the question is located.

Owing to this, the mini-model can have a simple mathematical form and its learning process can be considerably shorter than of the global model, e.g. of the global neural-network model. A mini-model can have both linear and non-linear form. In this paper the most simple linear-segment mini-models in the 2-D space will be presented. In [6] were presented triangle-linear mini-models. In 3D-space also tetragonal, pentagonal, and more complicated mini-models can be used. Figure 1 shows example of a triangle, linear, constrained mini-model, which has the task to calculate an answer for the following question:

*What does* $y^*$ *amount to if* $\{x_1^*, x_2^*\}=\{0.28, 0.88\}$?

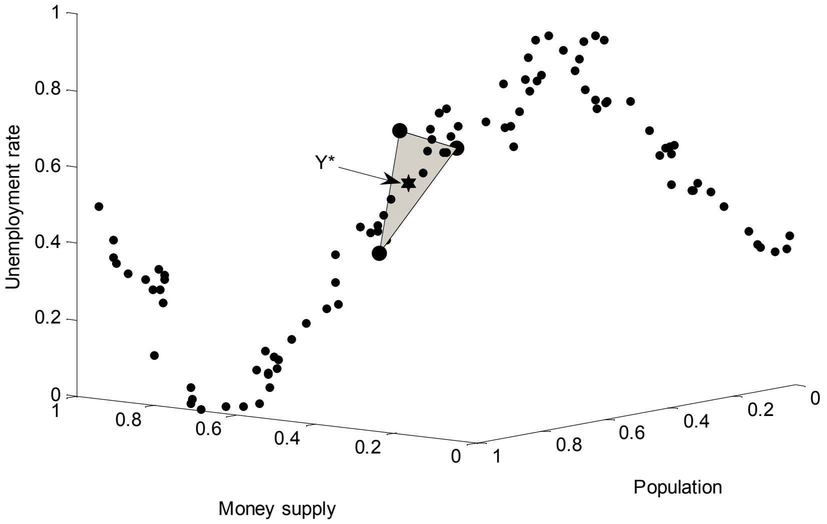


Figure 1. Example of a mini-model learned with samples to calculate the answer for the question "*What does* y *amounts to if* $\{x_1^*, x_2^*\}=\{0.28, 0.86\}$?". The answer calculated by the tuned mini-model is "$y^* = 0.621$". Source: own preparation.

The aim of this paper is not to explain the construction and the algorithm for training of mini-models with samples because of the paper-volume limitation, but to explain the difference between mini-models and the KNN-method of modeling. The aim is also to explain advantages of mini-models. Mini-models learn and calculate the numerical answer very quickly in comparison with global models. The reason of this fact is that they use only limited number of learning samples that lie in the neighborhood of the question point and also because mathematical form of mini-models is very simple. If new samples are achieved they can be quickly introduced to the data bank of the problem and used to easily actualize the previous mini-model and the answer of it. Global models encircle large spaces and are frequently strongly non-linear. Therefore their learning is much longer and sometimes not effective because of the local-minima phenomenon occurring in the learning process. Almost all experiments realized by the authors showed that mini-models are superior to such modeling methods as polynomials, neural network of the GRRN-type, and the KNN-method. Especially the comparison with the KNN-method is very important because the both methods applied for regression task calculate the answer for the question:

“*What does* $y^*$ *amounts to if* $\{x_1^*, x_2^*\}=\{..., ...\}$?”

on the basis of samples from the neighborhood of the question point $\{x_1^*, x_2^*\}$ where $x_1^*$ and $x_2^*$ are numerical values of the explanatory variables $x_1$ and $x_2$.
The KNN-method has been described in many books, e.g. in [1,3,5,8]. It determines the answer for a question on the basis of values of $y_i$, i=1,2,...,k of the k-nearest neighbors (samples) of the question $\{x_1^*, x_2^*\}$. However, the neighbor samples are multiplied by weights of contribution. A common weighting scheme is to give each neighbor a weight of l/d, where d is the distance of the question point $\{x_1^*, x_2^*\}$ to the neighbor in the space of the explanatory variables. The KNN-method is evaluated by many scientists as very effective and some of them are of the opinion that other methods are not necessary [5] (“Do we need whatever more than KNN?”). Experimental comparisons of the KNN-method and of the mini-model method have shown better accuracy of mini-models. But this superiority can also be explained essentially, what will be done in next chapters.

## 2. EXPERIMENTS OF MODELING THE DEPENDENCE y=f(x) IN 2D-SPACE (UNEMPLOYMENT RATE y IN POLAND AS A FUNCTION OF MONEY SUPPLY x)

The question, which factors influence the unemployment rate, is important and interesting for the government of any country. Investigations on the unemployment rate in Poland in years 1991-1999 were made in University of Szczecin and published in [7]. The investigations shown that the unemployment rate y depends on 7 factors x, which are ranked below from the most to the less significant one:
$x_1$ – money supply, $x_2$ – population, $x_3$ – dollar/zloty rate of exchange, $x_4$ – import value, $x_5$ – rediscount rate, $x_6$ – household expenditures, $x_7$ – number of high school graduates. In [6] the authors presented results of modeling the unemployment rate $y=f(x_1, x_2)$ in 3-D space. In this paper modeling of the dependence $y=f(x_1)$ will be shown, were x was the most significant and the strongest factor influencing the unemployment rate in Poland in years 1991-1999. The strength of this factor is testified by the fact that it delivers modeling errors of the order of few percents.

Below there are shown modeling errors that were achieved with various modeling methods:

a) modeling with polynomials $y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$ of order 5, the mean absolute error (MAE) = 0.0725
b) modeling with the GRNN-neural network (General Regression Neural Networks), MAE=0.047
c) modeling with the K-nearest neighbors (KNN) method, MAE= 0.093.
d) modeling with linear-segment mini-models, MAE=0.021.

Also in other experiments of modeling, which will not be presented here, mini-models achieved larger accuracy than other modeling methods. In the next chapter there will be presented 3 steps of the learning process of the mini-model in 2D-space to enable better understanding of this process.

## 3. ILLUSTRATIVE EXAMPLE OF THE MINI-MODEL LEARNING IN THE 2D-SPACE

The task of the mini-model in the presented example was to calculate the answer for the question:

*What does the unemployment rate $y^*$ amounts to if the money supply $x^* = 0,28$?*

To calculate the possibly credible answer the mini-model has at first to tune itself on the basis of samples of the dependence y = f(x), which encircle the question point QP = {0.28, ?} in the X-space, Figure 2.

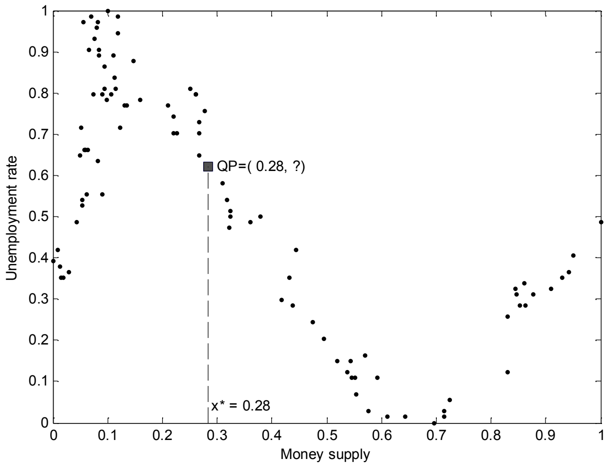
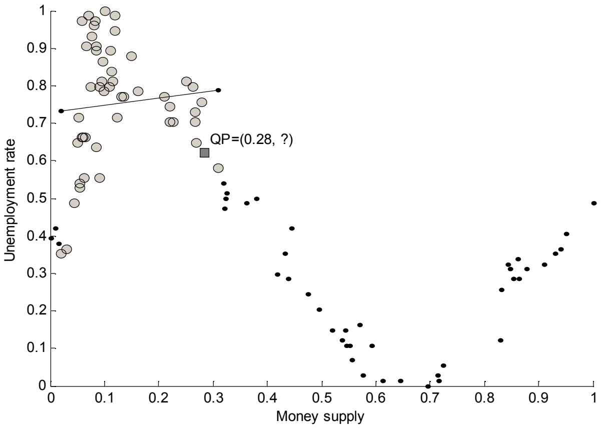


Figure 2. Example of a mini-model learned with samples to calculate the answer for the question "*What does* $y^*$ *amounts to if* $x^* = 0.28$?. Source: own preparation.

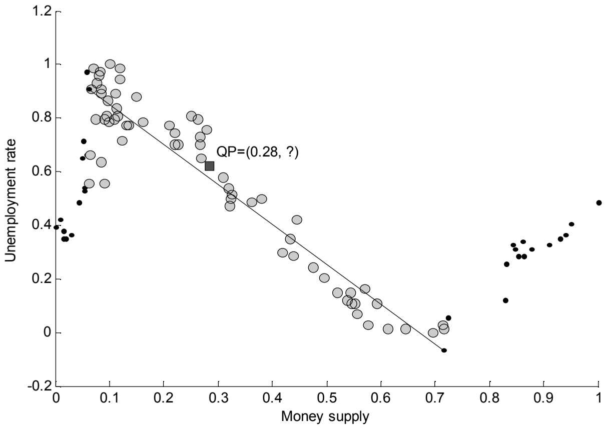
In the learning process with samples the mini-model have to find its optimal length, gradient (inclination) and position in the X∗Y-space. These are 3 factors that have to be determined. In the case of the KNN-method only one factor (position) is determined. Only when all the 3 factors have been determined the mini-model calculates the answer $y^*$ for the question point $x^*$ .

In Figures 3a, 3b, 3c three steps chosen from many steps of the learning process in 2D-space were shown. Figure 3a presents the start position of the linear-segment mini-model. The start position, inclination and length of the segment was chosen by the computer program by chance. The only condition which has to be satisfied is that the question point ($x^*$), which in this case is {0.28} has always be encircled by the mini-model range.

a)



b)



c)

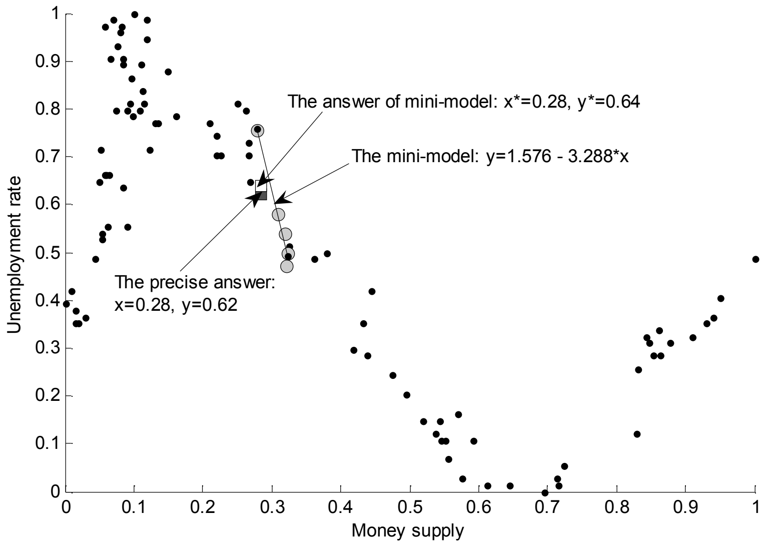
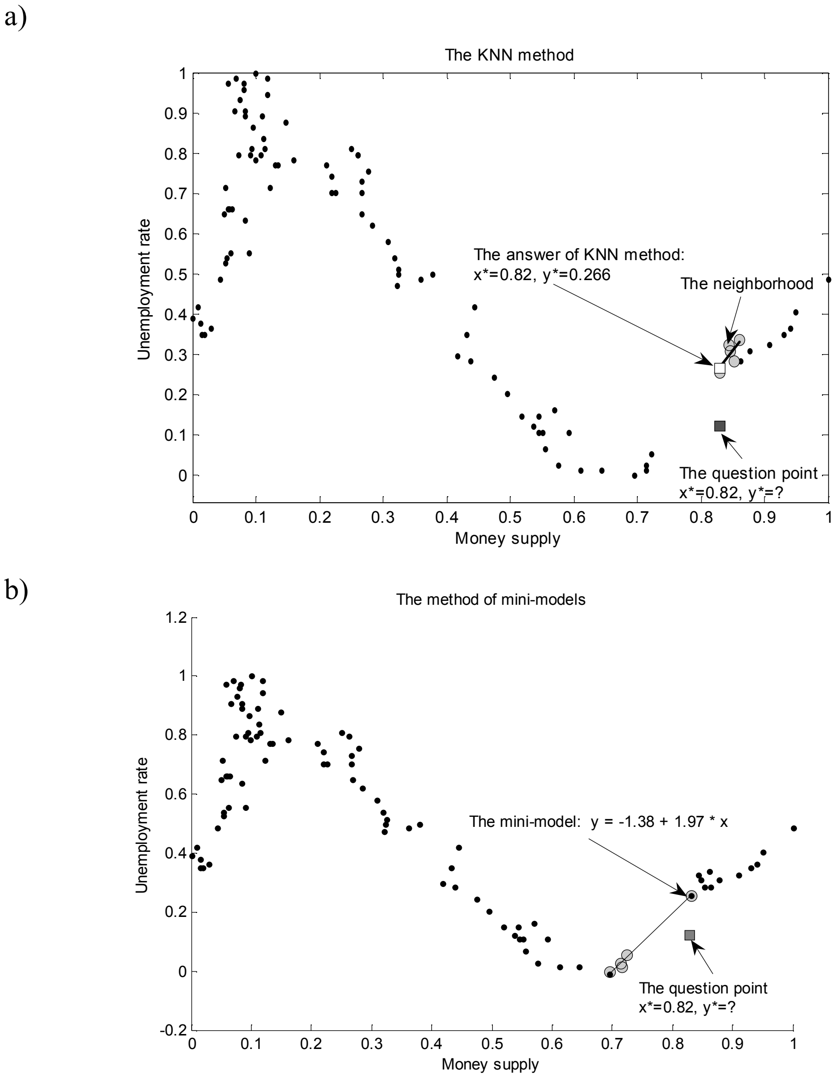


Figure 3a. The start position of the mini-model which has to calculate the answer for the question *"What does y amounts to if* $x^*$ *= 0.28?"*. Figure 3b. One of the intermediate positions of the mini-models assumed by it in the learning process. Figure 3c. The end, optimal position of the mini-models, y=1.58–3.29x, QP = [0.28, 0.64]. Source: own preparation.

After finishing the learning process the mini-model calculated the answer $y^*$= 0,64. The precise and correct answer was in this case known (y = 0,62) because the question referred to one of the samples, which was taken out of the learning process (the sample gave values of unemployment rate and money supply from July 1995). The method of mini-models also was tested with the known cross-validation method "leave one out" [2] for each of the 96 samples representing particular months of the period 1991-1999. The mean absolute error (MAE) was equal to 0.021 at the variables x and y normalized to the interval [0, 1].
In the next chapter the difference between the method of mini-models and of KNN will be discussed.

## 4. DIFFERENCES BETWEEN MINI-MODELS AND THE K-NEAREST NEIGHBOR METHOD

The method of mini-models calculates the numerical value of the explained variable $y^*$ on the basis of samples, which have the x-value positioned near to the x-value of the question. For this reason the Reader may mistake this method as a method being identical with the KNN-one, that is commonly known in the scientific world, good tested and verified. Therefore using mini-models could appear nonsensical. Such opinion would be highly erroneous because, as it will be shown, the mini-model method is a much more refined one and in most cases delivers more precise calculation results than the KNN method. To enable the Reader better understanding mini-models specially the 2-dimensional dependence y=f(x) has been chosen. In Figure 4a the example question *"What does y amounts to if* $x^* = 0.82$*?* and the answer $y^*$ calculated by the KNN-method is presented.



c)

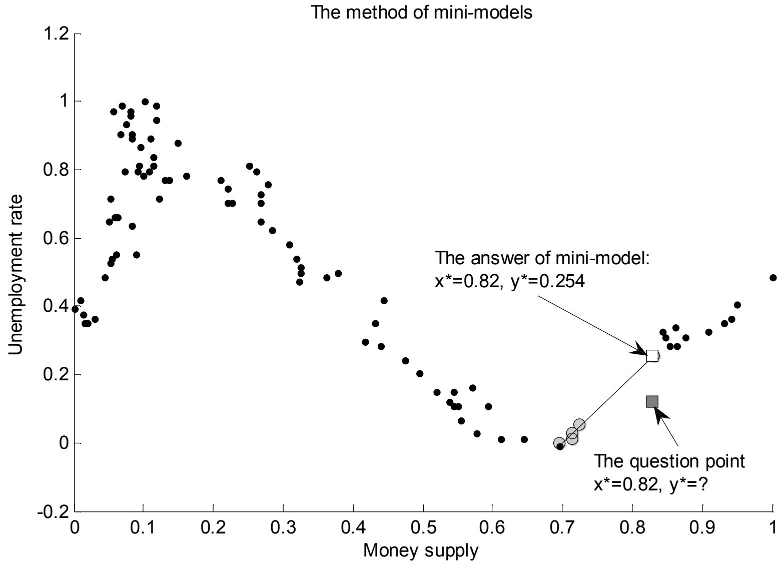


Figure 4. Illustration of differences between the KNN (figure 4a) and the mini-model method (figure 4b, 4c) in calculating the predicted y-value in the case of existence of information gaps in learning samples. Source: own preparation.

The question point $x^{*}$= 0.82 presented in Figure 4 is positioned in an information gap, in the range with local lack of learning samples. However, the point is positioned nearer the right side than the left side of the gap. In this situation the KNN-method calculates the answer $y^{*}$ for the question point QP=[0.28, ?] mainly on the basis of the nearest samples, which are situated on the right side of the question. The calculated answer (prediction) is $y^{*}$ = 0.266. As can be seen the KNN-method does not take under consideration (or takes very weakly) the learning samples that lie on the left side of the question point, on the left side of the information gap. However, the mini-model method does it, Figure 4b. Therefore its result, taking only human intuition into account, is more credible.

This credibility and accuracy of mini-models can not only intuitively but also objectively be verified with the cross-validation method "leave one out" [2] when for the query point one of samples is chosen and query is positioned among $x^{*}$ samples. It this case accurate answer $y^{*}$ should be equal to the position of y sample. For testing both methods a sample representing December 1994 was chosen. Both methods were given the task to calculate answer for question point equal to the x-value of the sample taken out. The results achieved are shown

below.
The tested sample: x (money supply) = 0.25, y (unemployment rate) = 0.81.
The question: *"What does $y^*$ amounts to if $x^*$ = 0.25?"*
The KNN-method: the answer $y^*$ = 1.014, the mean absolute error MAE = 0.20296
The mini-model method: the answer $y^*$ = 0.774, MAE = 0.03705.

The above example shows that the KNN-method is not effective regression method in information gaps contrary to mini-models. If the question point lies nearer to one border of an information gap the KNN-method chooses for calculation of the answer y samples of the nearest border. Thus, this method loses information about the tendency of changes dy/dx of the modeled dependence. And the mini-model method identifies this tendency and takes it into account at calculating the answer $y^*$.

Additionally, the length of the mini-model is the third information identified by the mini-model (information about the validity range of the mini-model) apart from its position and inclination dy/dx. Thus, the information delivered by the mini-model is much richer than that delivered by the KNN-method.

Additionally, the superiority of mini-models is supported by the fact that in the cross-validation test for all 96 samples (not only for one sample as in the example above) the mini-model method has shown in most experiments higher prediction precision than the KNN-method. In the experiment described in this paper the mean absolute error MAE equals for the mini-model 0.021 whereas the KNN method had the error equal to 0.093 for the unemployment rate y normalized to the interval [0,1].

On example of samples shown in Figure 1 one could think that existence of information gaps in samples is a rare event. Unfortunately, the truth is the other way round. In the case of dependence y=f(x) investigated in this paper, 96 samples seems to be a considerable number, which tightly covers the interval [0, 1] of variable x (normalized money supply). However, when the unemployment rate is modeled in 3D space, $y=f(x_1, x_2)$, where $x_1$ – money supply, $x_2$ – population number, the same 96 samples have to cover the area [0, 1]*[0, 1] of the input space $X_1*X_2$. This number of samples covers input space only fragmentary and a larger part of this space is without samples and creates great information gap. Then no questions concerning the unemployment rate are allowed in this space part. Participation of the space without any samples in the full input space drastically increases with the number of explanatory variables x taken into account in modeling, e.g. when the unemployment rate is modeled in 4D or larger space. As investigations described in [4] have shown modeling multidimensional dependencies almost always is connected with considerable information gaps. For this reason the mini-model method increases its superiority over the KNN-method owing to its capability of identifying tendencies dy/dx in samples and of the mini-model-lengths.

## 5. CONCLUSION

The mini-model method is an effective method of modeling that gives good prediction results in regression tasks. The authors have not tested classification applications of this method until now. Mini-models have some similarity with the popular KNN-method and can easily be mistaken with it. The paper presented main differences between both methods and also explained why mini-models are more precise than models delivered by the KNN-method. The authors have investigated modeling capabilities of mini-models in 2D- and 3D-space and at present they are preparing computer programs and experiments with this method for higher spaces.

### *REFERENCES*

[1] Cherkassky V., Mulier F. (2007) *Learning from data,* Wiley-Interscience, IEEE Press.

[2] Geissler S. (1993) *Predictive inference.* Chapman and Hall, New York.

[3] Hand D., Mannila H., Smyth P. (2001) *Principles of data mining.* Massachusetts Institute of Technology, USA.

[4] Klęsk P. (2008) *Construction of a neurofuzzy network capable of extrapolating (and interpolating) with respect to the convex hull of a set of input samples in Rn.* IEEE Transactions on Fuzzy Systems, vol 16, issue 5, pp.1161-1179.

[5] Kordos M., Blachnik M., Strzempa D. (2010) *Do we need whatever more than k-NN*? Proceedings of 10-th International Conference on Artificial Intelligence and Soft Computing, Zakopane, Poland, June 13-17, 2010, Springer, Germany, pp.414-421.

[6] Piegat A., Wąsikowska B., Korzeń M. (2010) *Zastosowania samouczącego się, 3-punktowego minimodelu do modelowania stopy bezrobocia w Polsce.* Artykuł ukaże się w 2010 roku w zeszycie naukowym Studia Informatica Uniwersytetu Szczecińskiego.

[7] Rejer I. (2003) *Metody modelowania wielkowymiarowego systemu z użyciem metod sztucznej inteligencji na przykładzie bezrobocia w Polsce.* Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

[8] Witten I.A., Frank E. (2005) *Data mining.* Morgan Kaufmann Publishers, San Francisco.